ConTour: Data-Driven Exploration of Multi-Relational Datasets for Drug Discovery

ConTour: Data-Driven Exploration of Multi-Relational Datasets for Drug Discovery

Christian Partl, Alexander Lex, Marc Streit, Hendrik Strobelt,
Anne-Mai Wassermann, Hanspeter Pfister and Dieter Schmalstieg

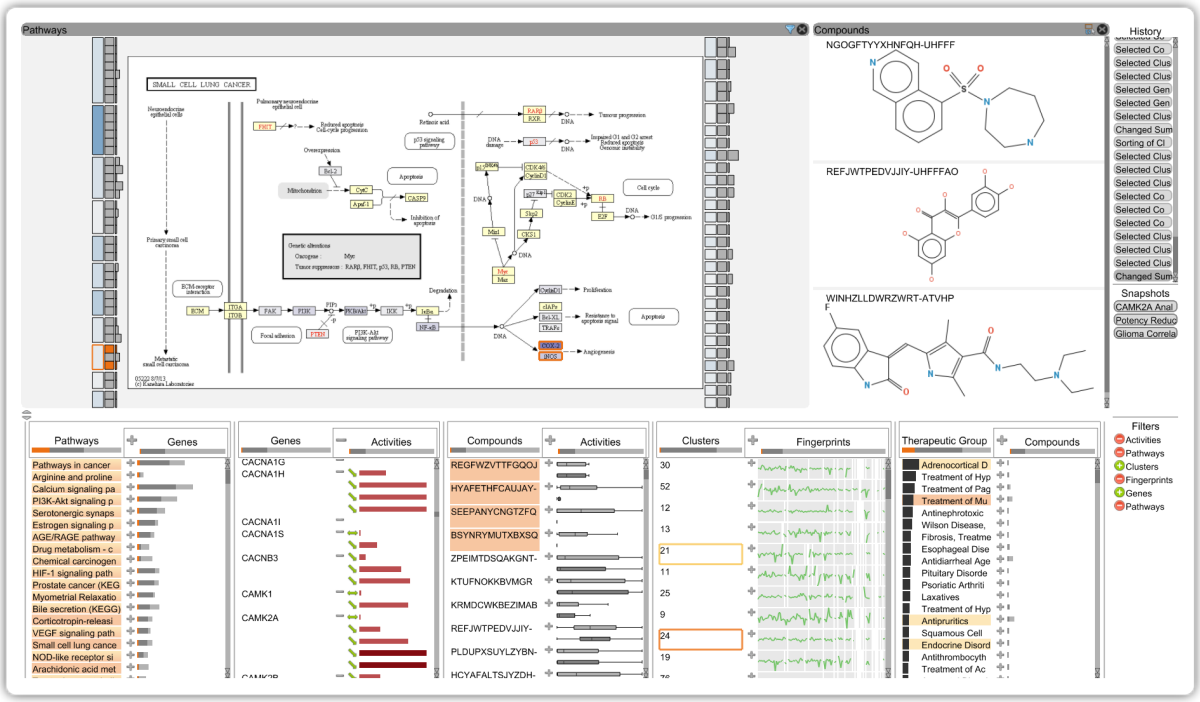


Fig. 1. ConTour shows a multitude of heterogeneous data items in several columns in the relationship view (bottom). The detail views display a selected pathway and selected chemical structures of compounds (top).

Abstract—Large scale data analysis is nowadays a crucial part of drug discovery. Biologists and chemists need to quickly explore and evaluate potentially effective yet safe compounds based on many datasets that are in relationship with each other. However, there is a lack of tools that support them in these processes. To remedy this, we developed ConTour, an interactive visual analytics technique that enables the exploration of these complex, multi-relational datasets. At its core ConTour lists all items of each dataset in a column. Relationships between the columns are revealed through interaction: selecting one or multiple items in one column highlights and re-sorts the items in other columns. Filters based on relationships enable drilling down into the large data space. To identify interesting items in the first place, ConTour employs advanced sorting strategies, including strategies based on connectivity strength and uniqueness, as well as sorting based on item attributes. ConTour also introduces interactive nesting of columns, a powerful method to show the related items of a child column for each item in the parent column. Within the columns, ConTour shows rich attribute data about the items as well as information about the connection strengths to other datasets. Finally, ConTour provides a number of detail views, which can show items from multiple datasets and their associated data at the same time. We demonstrate the utility of our system in case studies conducted with a team of chemical biologists, who investigate the effects of chemical compounds on cells and need to understand the underlying mechanisms.

Index Terms—Multi-relational data, visual data analysis, drug discovery

- *Christian Partl and Dieter Schmalstieg are with Graz University of Technology. E-mail: {partl, schmalstieg}@icg.tugraz.at.*
- *Alexander Lex, Hendrik Strobelt and Hanspeter Pfister are with Harvard University. E-mail: {alex, strobelt, pfister}@seas.harvard.edu.*
- *Marc Streit is with Johannes Kepler University Linz. E-mail: marc.streit@jku.at.*
- *Anne-Mai Wassermann is with Novartis Institutes for BioMedical Research. E-mail: anne_mai.wassermann@novartis.com.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

1 Introduction

The need to explore multi-relational data is common in many domains. Answering questions such as whether relationships of particular entities across datasets exist or how strong or specific a relationship is, is important for a variety of applications. This is also true in drug discovery. Researchers want to learn whether there are chemical compounds, i.e., drugs or drug candidates, that modulate a specific biological process without influencing others, or want to see which drugs induce a characteristic change in a cell's phenotype. However, due to the complexity of the experimental data, the manifold interactions between compounds and cellular components, and the rich associated data, the

analysis is difficult and cannot efficiently be done with current tools.

The fundamental challenge in such an analysis is resolving the web of relationships while providing the relevant associated data. Which drug inhibits the relevant biological function at the right concentration and exhibits an actual effect on the organism? Answering this and similar questions requires resolving relationships between multiple datasets, ranking results, filtering based on attributes, and visualizing the context of the involved components.

To address this challenge, we developed ConTour, a novel visual analysis tool that supports researchers in drug discovery. ConTour employs a drill down approach to untangle the web of relationships: analysts select entries in one of the datasets they are interested in and ConTour presents the related entries. To support the analytical tasks of chemical biologists in drug discovery, we introduce a novel visualization concept for flexibly arranging and nesting datasets, which makes the browsing of relationships more user-friendly. ConTour employs multiple advanced visualization methods on different levels of detail for the analysis of data items and relationships. By providing parameterizable enrichment scores, ConTour allows analysts to investigate complex relationships that involve multiple different data types.

We demonstrate the applicability of ConTour for its designed purpose in three case studies, which reveal relationships between various types of biological and pharmaceutical data, including pathways, genes, compounds, compound activities, biological fingerprints, therapeutic groups, and clusters. Using ConTour, our collaborators detected correlations between fingerprint clusters and therapeutic groups and were able to explain the composition of fingerprint clusters by common targets in pathways. These findings are valuable indicators in support of our collaborators' hypothesis that biological fingerprints can be used to detect meaningful compound similarities and that fingerprints also reflect the effect of compounds on the cell or organism.

The drug discovery domain problem can be generalized to the problem of analyzing multi-relational datasets. The relationships between genes, pathways, drugs, and further entities are, from a computer science perspective, similar to those, for example, between customers, vendors, and products in relational databases. Consequently, we argue that our approach is applicable to many other problems.

2 Related work

The multi-relational data exploration problem can be interpreted as a graph exploration problem where each item of each dataset represents a node and the relationships between the items are the edges. ConTour also employs a faceted search approach, where the considered data is continuously narrowed down, either based on explicit data relationships or based on attribute filters. Consequently, we review the literature regarding both topics, in addition to a discussion of relevant visualization approaches in pharmacology and molecular biology.

Graph exploration Jigsaw's list view [24, 7] visualizes relationships between multiple entities. Each list in the list view can be understood as a partition of a graph. Selecting an item highlights the related items in the other lists, facilitating a query-driven analysis. Schulz et al. describe a similar table-based approach to visualizing bi-partite biological networks [19]. In contrast to Jigsaw's list view, each partition of the network is drawn in a table, which can be sorted based on various attributes. The two partitions are connected with links. Both tools visualize attributes within the cells. Ghani et al. [6] conducted a design study on multimodal social network analysis and developed *parallel node-link bands*, which are also similar to Jigsaw's list view. Social network data exhibits similar characteristics to the data discussed in this paper. Ghani et al.'s evaluation shows that the parallel division of items into multiple columns was easy to understand and worked well for the domain experts. All of these interfaces use visible links to associate the individual columns, which is useful if all list items can be fit on the screen, but less helpful when the targets of links are not in the viewport. None of these approaches enable nested embedding of partitions within each other.

GraphTrail [4] uses interaction to navigate a large and heterogeneous network using multiple charts. New charts can be duplicates of their parent, or can be a refinement of the data relative to the first one. This refinement can be done using the nodes directly or based on attributes. While GraphTrail does not explicitly represent the network partitions, as ConTour does, we use a similar approach to continuously refine a selection to drill down into the dataset.

Lieberman et al. [12] employ the semantic substrates concept [21] of visualizing connections between different semantic partitions of a network to relationships across biomedical datasets, between, for example genes to PubMed and OMIM (a disease database). It represents the items of each dataset or class in a scatterplot and draws explicit links between them, which, however does not allow to embed rich meta data in the nodes.

Faceted browsing InfoZoom [22] and FOCUS [23] are examples of early faceted browsing systems that allow a linear drill down and provide focus and context technique similar to Table Lens [18]. Yee et al. [29] describe a method to search for images along conceptual dimensions, which is similar to our approach of refining queries based on selections in multiple datasets. PivotPath [3] introduces an informal approach to interacting with faceted datasets on the web by explicitly presenting connections between various facets, which, combined with animated transitions between filters, makes the results of queries more comprehensive. The system is conceived to encourage exploration and lacks a data-driven approach, such as filtering based on attributes. PivotSlice [30] visualizes both explicit and implicit relationships in a citation network and allows multi-focus exploration.

ConTour is related to all these approaches in the way it enables drill down into a complex dataset, yet also distinct since the facets in ConTour also correspond to the query results. Furthermore, none of these approaches use network-based or other metrics to rank and filter the items in facets.

Pharmacology visualization The visual analysis model to drug candidate selection by Konecni et al. [10] can be considered complementary to ours. It uses machine learning to select compounds from a large library and visualization to evaluate and update the model.

Becker [2] visualizes structural similarities of compounds by mapping the compound attributes to the axes of parallel coordinates plots. The approach, however, does not integrate other data sources.

In contrast, Lounkine et al. [13] also considers the interaction of compounds with biochemical pathways. They classify compounds based on their structure and visualize their interactions with pathway nodes. Since many compounds interact with specific nodes, however, the results can be cluttered. A design goal of ConTour is to reduce the number of compounds to those that are truly relevant and thus avoid similar problems.

enRoute [15] takes an alternative approach to pathway visualization to address the scalability problem of plotting rich attribute data on top of nodes in a graph. It enables analysts to select individual paths, which are extracted and for which detailed information is visualized. For compound-pathway interactions, however, the data along a specific path is less relevant than the topology of the network and the various interaction partners of a specific class of compounds.

HiTSEE [27] helps finding correlations between the structure of chemical compounds and their activity in reactions with a biological target. Compounds can be selected and projected with respect to structural similarity. Starting with a seed set, the user can expand the selection involving neighbors. Clusters of compounds show common substructures to allow reasoning about which molecule substructures are driving activity. HiTSEE focuses on an in-depth analysis of one to one relationships between structures and activities, whereas ConTour takes a broader view and aims to identify potentially relevant relationships.

3 Biological background and domain goals

For many years, drug discovery has focused on finding the "magic bullet", i.e., the identification of a drug that selectively interacts with a disease-causing or pathology-relevant protein target [26]. However, with more and more data describing how drugs interact with biomolecules (bioactivity) and a better understanding of the biological network, evidence accumulates that this strategy employs an overly

simplistic view of human disease and drug-target relationships. Indeed, existing bioactivity data suggests that approved drugs interact on average with seven different protein targets [14]. Furthermore, one protein target can be involved in many different biological processes. Therefore, its modulation by a compound can influence multiple, seemingly unrelated phenotypic traits, i.e., have multiple observable effects on the organism, both on a cellular and whole organism level. However, the same phenotype can be induced by compounds that interact with different protein targets, e.g., if the proteins are part of the same signaling pathway. A pathway is a meaningful set of biomolecules and reactions, whose interplay fulfills a particular function in a cell or organism. Given this complexity of the biological system, classical structure-activity relationship analyses that study the effect of a compound set against one particular protein need to be complemented by techniques that allow for a more holistic view on the effects that a compound has on a biological network. In the pharmaceutical industry, historical experimental data can be leveraged and combined to generate so-called *biological fingerprints* that report the activity of a compound across dozens of experiments that were designed to monitor different cellular processes. In essence, the fingerprints describe a numerical characterization of different experimentally measured phenotypes. They are thus numerical descriptions of the observable effect of a drug on a cell or organism and provide a more comprehensive view on the manifold biological actions of a compound than simple protein-compound interaction data.

Comparison and clustering of compounds based on the biological fingerprints can lead to the detection of novel compound-target or compound-disease relationships. More specifically, our collaborators have three analysis goals:

- **Identify a drug's mechanism of action.** If a compound with an unknown mechanism-of-action falls into a cluster where all other cluster members are known to modulate the same protein target, it is conceivable that the compound also binds to this target.
- **Identify the biological process a drug modulates.** If compounds that bind to different targets cluster tightly together, one can hypothesize that these targets are involved in the same biological process.
- **Identify new drugs for specific therapeutic indications.** A compound that clusters together with drugs for a particular therapeutic indication could be a novel candidate drug for this therapy, with potentially advantageous properties.

While the biological fingerprints are the foundation of such an analysis, multiple other datasets also need to be considered to paint a holistic picture, all of which we describe in the following section.

4 Dataset description

The drug dataset studied in this paper consists of about 1,100 compounds that have been extracted from the public bioactivity databases *ChEMBL* [5] and *DrugBank* [11]. Over the past decades, all of these compounds have been profiled in at least 50 different cell-based screens at the pharmaceutical company Novartis [16]. These screens were tested for compound activities against a panel of diverse targets, pathways, and organisms. In each screen, all compounds were tested at a single concentration, and compound activities were reported in form of Z-scores, i.e., the number of standard deviations that a compound's effect in a screen differed from the mean response of all compounds tested in the screen. For each compound under study, its Z-scores were combined into a vector (the "fingerprint"), where each position was associated with a specific assay. Overall, Z-scores from 105 different assays were considered in the generation of the compound fingerprints. A correlation-based similarity measure [28] was used to calculate a similarity matrix between all compounds, which was then used as input for hierarchical clustering. The resulting dendrogram was divided into 100 distinct clusters. These fingerprint clusters provided the basis for our analysis.

The compounds in the dataset were annotated with meta-data. First of all, all compounds were annotated with about 7,000 activities against protein targets that were extracted from the data repositories ChEMBL and GVK[1]. These activities describe whether there is a known interaction for a particular compound-protein pair. Compounds typically interact with multiple proteins, and proteins also interact with multiple compounds. The dataset distinguishes between three types of interactions: target activation, inhibition, and binding. That means we discriminate between compounds that in- or decrease the activity (functional effect) of a protein. If, based on the reported bioactivity data, the direction of the interaction cannot be inferred, it is reported as a binding event. In addition, these activities reported AC50 values, i.e., the concentration of the drug at which 50% of the maximal response was achieved. For example, for a compound that inhibits an enzyme that cleaves other proteins, the AC50 value is the compound concentration at which the observed cleavage is reduced by 50%. AC50 values thus characterize the potency of a drug; the lower the value, the more potent the drug.

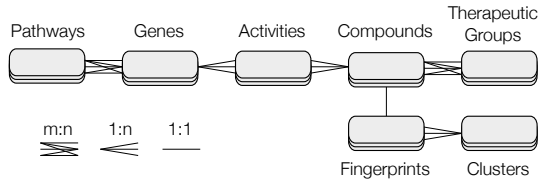


Fig. 2. Overview of pharmaceutical and biological data sets that are analyzed using ConTour. The edges of this graph indicate direct relationships between data items and the cardinality of these relationships. A pathway contains multiple genes, and one gene can be part of multiple pathways. Activities connect genes and compounds. Compounds represent drugs, drug candidates, or other small molecules and are classified into therapeutic groups. Fingerprints describe biological properties of compounds. Fingerprint clusters can reveal biologically relevant similarities of compounds.

The about 1,100 considered proteins were mapped to their corresponding genes; we will use either terms hereafter. The biological roles of proteins in the human organism are captured in about 450 *KEGG* [8] and *Wiki* Pathways [9]. Many of the compounds in the studied dataset are approved drugs or clinical candidates. They were classified into about 400 therapeutic groups using the classification scheme of the *Prous Integrity database*[2]. Figure 2 provides an overview of all involved data and their relationships.

The structure of the available data can be described as a k-partite graph, where sets of items, such as pathways, genes, compounds etc., represent the partitions of the graph. This implies that the items within a set have no defined relationship, but that relationships are defined between items of different sets. The graph describing the set relationships (see Figure 2) is connected and acyclic, i.e., there are no sets that are not related to others, and there is exactly one path connecting any two sets. The relationships between the items of the sets can be of arbitrary cardinality (1:1, 1:n, or n:m). Though this graph only shows *direct* relationships, ConTour is designed to also consider *indirect* relationships via intermediate sets and items. Pathways, for example, are indirectly connected to compounds via genes and activities. Thus, we refer to items as being *related* when they are directly or indirectly related.

5 Task analysis

In repeated consultations with multiple domain experts over half a year we elicited a set of tasks an analyst has to perform to achieve the previously described domain goals. It is worth noting that none of the domain goals can be directly mapped to a specific set or sequence of tasks, but rather, that all of these tasks have to be executed in an iterative, open analysis session to reach any of the domain goals. These tasks are:

[1] http://www.gostardb.com/

[2] https://integrity.thomson-pharma.com/integrity/xmlxsl/

T 1: Identify related items. Given an item of type A, find all items of type B that are directly or indirectly related. An example of this task is to identify all pathways that contain a specific gene (direct), or all compounds that influence a pathway (indirect).

T 2: Identify items that share a relationship with a set of items. Given a set of items, find all items that are connected to all of the input items. In other words, identify the items that all of the input items are related to. The input items can be from the same set or from different sets. An example is to identify all genes that are shared between two pathways, or to identify all compounds that are connected to a specific cluster and that are also related to a specific pathway.

T 3: Analyze network enrichment. In highly relational datasets many nodes are connected, directly or indirectly, to many others, which can lead to unspecific relationships. Our collaborators, however, are interested to identify the connections that are very specific. For example, they want to identify clusters of compounds where all compounds interact with only one specific pathway. More generally, for items of type A and B that are not directly related, one might want to judge how closely they are connected by considering items in the chain between them.

T 4: Rank items. Being able to rank items is crucial to reveal the most important items out of a long list. Rankings can be based on item attributes or on derived measures such as network enrichment.

T 5: Filter items. Analysts want to filter items, either based on attribute values or based on relationships. An example for the former is that an analyst might want to only consider activities that activate their interaction partner, and ignore inhibiting or binding drugs. The latter case depends on T 1 and T 2 - items that are not related to a specific selection of items should be filtered out.

T 6: View items in detail. The relevance of data items can often only be judged by exploring their attributes. While some items are simple, such as activities, others, such as pathways or compounds, are complex entities. A central task is to view these complex entities in detail. For example, a pathway should be viewable in all its complexity, or the chemical structure of compounds should be displayable.

6 Concept

The previously introduced tasks describe an analysis process that is highly exploratory in nature, rather than a rigid step-by-step process with well-defined starting, intermediate, and end points. To enable such an analysis for different item sets, a visual analysis technique needs to allow analysts to flexibly gain access to information encoded by items or item relationships at virtually any point during the analysis.

Our approach to this problem is illustrated in Figure 3. The **data graph** component contains all data items, their relationships, as well as associated data present in the system. The items of this graph are presented to the user in the **visual interface**. The main component of this visual interface is the *relationship view*, which consists of a collection of columns, each listing the item set of a particular type. A second important component of the visual interface are *detail views*, which display detailed item information using representations specifically tailored to the item type. Based on individual items or whole item sets, several operations such as selecting, filtering, or nesting can interactively be triggered from the visual interface. Using graph information of the data structure, these operations are propagated to related items in other item sets, updating their representations in the relationship view and the detail views, e.g., by highlighting, hiding, showing, or reordering items. The tight interplay of the data graph and the visual interface realizes a highly interactive data-driven exploration of item relationships. In the following sections, we will discuss the components of the visual interface in more detail.

6.1 Relationship view

The relationship view is composed of several freely arrangeable columns that represent one item set each. Individual columns can be scrolled, sorted, and filtered independently. The layout is designed to enable arbitrary entry points into the analysis: every item in every column can be a starting point. The column's header displays relevant

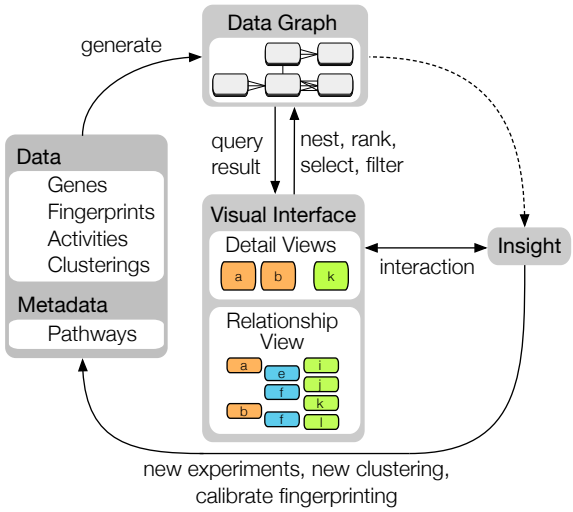


Fig. 3. The overall visual analytics process in ConTour. Data from internal or external (public) sources make up the data graph. The data graph is the underlying data structure for the visual interface and responds to its queries (e.g., nesting, ranking, etc.). Through interaction with the visual interface domain experts generate insights. These insights can be of value on their own, or can lead to refined biochemical experiments, new ideas for clusterings and groupings, or to calibrated fingerprint algorithms, thus generating new data.

summary information, while the body displays the items. Items show either a label, or relevant attribute information in built-in visualizations, or both. Columns can be added, duplicated, and removed at any time. As columns can potentially contain thousands of items, we provide several operations that can be performed on individual items or whole columns to explore this huge amount of items and relationships effectively. These operations are discussed in the following.

Item selection and highlighting. One simple yet effective method to find related items (T 1) is highlighting. Selecting an item highlights all of its related items. We distinguish between two selection methods: Hovering over an item just highlights all related items, whereas clicking on an item also moves all related items in all columns to the top. As this reordering might be undesirable in some cases, for instance, if the items of a column use a meaningful sorting, it can be disabled. When selecting multiple items of a type, we employ one of two different modes, which are illustrated in Figure 4, to combine the highlights: In **union mode**, all items that are related to any selected item are highlighted. In **intersection mode**, only those items that are related to all selected items are highlighted.

Selection-based filters. Selection-based filters allow to reduce the whole data space to those items that are related to selected items (T 1, T 5). Applying multiple filters in succession gradually narrows down the data space. In essence, each newly applied filter is combined with the result of all previous filters using a Boolean `and` operation. However, after the data space is narrowed down, it might be desirable to expand it again. Therefore, we provide the possibility to add related

Fig. 4. Illustration of the two highlight modes. In union mode, selecting items *a* and *b* highlights all items related to either of them, i.e., items *i*, *j*, and *k*. In intersection mode, only items related to both *a* and *b* are highlighted, which is item *j* in this case.

items that have previously been filtered out. This additional operation can be regarded as a filter that is combined with the result of all previous filters using a Boolean `or`. The different selection modes (union, intersection) affect the filtering behavior. The data space gets reduced to the union or intersection of related items, respectively, and the union or intersection of related items are added.

Nesting. Nesting is an effective method to directly associate multiple related items of different sets. Columns can be nested to create parent-child relationships. Nesting two columns has the effect that for each item in the parent column, all related items of the child column are shown right next to it, as shown in Figure 5(a). Nesting a gene column within the pathway column, for example, displays all genes that a pathway contains next to the pathways. In contrast to highlighting, nesting always unambiguously shows what items are related even if multiple items are selected. The downside of nesting is that it is less space efficient as it results in redundant items. To remedy this, children can be collapsed, so that the relationships are shown only on demand. When child items are collapsed, a summary representation allows the analyst to gain an overview of these items, as shown on the right of Figure 5(a). This representation can show a simple count of the children or summary statistics about the children's attributes.

An interesting possibility is to use summary values of the children to sort the parent, opening up new opportunities to identify relevant items. To easily identify items with many relationships, for instance, the number of child items can be used as sorting criterion.

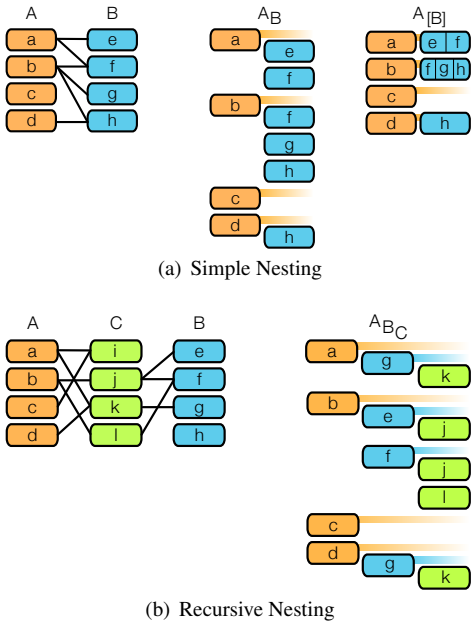
Nesting is a powerful way to investigate direct and indirect relationships between different items (T 1), as items of a child column may be intuitively associated with their parents for multiple items of the parent column at once. ConTour also allows to nest multiple columns. Thus, child columns can be siblings or be nested recursively, as illustrated in Figure 5(b). However, a recursively nested item is only considered as a child, if it is related to all of its parent items in the chain of parent columns. This makes recursive nesting equivalent to a filter chain applied to the items of the nested columns, with the filters being defined over the relationships to the parent items. Recursive nesting of columns is an effective way for identifying items that are commonly related among items from different sets (T 2). For example, in Figure 5(b) on the right, item *b* and *f* do have *j* and *l* in common.

Ranking and sorting. Ranking and sorting items in a column (T 4) is a simple method to identify the most interesting items quickly. The sorting criteria can be manifold. For example, items can be sorted alphabetically or by some numerical attribute. Rankings can also be based on scores (T 3), that quantify certain network properties. Sorted items can easily be compared, if their representation reflects the sorting criterion, such as attribute values.

Column-based filters. Applying filters to columns (T 5) is a simple method to remove uninteresting or irrelevant items. Similar to sorting, filters can be based on several criteria. For instance, attribute-based filters may define the value range for numerical attributes or filter items based on associated categories. A simple example is to remove all activities that are above a threshold in their AC50 values, which indicates that they are not potent. Filters may be applied locally or globally. **Local filters** only affect the item set of their column. **Global filters** affect all item sets by removing the items that are not connected to one of the remaining items in the source column. Global filters are efficient at reducing the complexity of the whole data space.

6.2 Detail views

Triggered from the relationship view, detail views show one or several items using suitable visualizations (T 6). Detail views are tailored to item types. Some may show all items, some a subset of items, and others may show only one item at a time. For example, the detail view for fingerprints shows all of them in one large parallel coordinates view. An example for a detail view that shows only selected items is the compound view, where only the selected compound structures are shown. Detail views can also integrate multiple item types. Our pathway view, for example, shows a pathway together with genes, compounds, and fingerprint clusters.



(a) Simple Nesting

(b) Recursive Nesting

Fig. 5. Nesting. (a) On the left direct relationships between the items of columns A and B are indicated as connecting lines. In the center, column B is nested in column A, placing related child items of column B right next to their parent item of column A. For example, items *e* and *f* are related to item *a* and are therefore displayed as its children. As item *f* is also related to *b* it is shown next to both parents. On the right, the child items of *a* and *b* are collapsed into summary representations. (b) On the left the direct relationships between items of column A and C and columns B and C are displayed. The items of columns A and B are indirectly related via items of column C. On the right the columns are recursively nested. Column B is nested in A, and C is nested in B. This recursive nesting helps to find items in C that are commonly related among items in A and B. For example, items *a* and *g* are commonly related to item *k*, whereas items *b* and *f* are related to items *j* and *l*.

7 Realization

We developed the prototype of ConTour in close cooperation with our collaborators, who gave feedback on a weekly basis. In this section, we describe the design decisions we made to represent the data, which algorithms we implemented to satisfy the analytical needs of our collaborators, and what additional tools we added to support the analysts in the data exploration process.

7.1 Relationship view

How we represent the various types of data items in the relationship view mainly depends on the amount of information held by each item. If an item has no additional data associated, we display its name or ID, which is the case for genes, clusters, and therapeutic groups. Also, if there is too much information available to fit in the columns, like in the case of pathways and compounds, we also only show their names or IDs. Although fingerprint items come with over 100 numerical values, it is still possible to visualize them in a compact way. To achieve this, we use centered bar charts with bars pointing up and down, as shown in Figure 6(d). This compact representation provides a good overview of the parameters and can be effectively used for comparative tasks. Fingerprints contain many missing values, which we encode by leaving the area empty, while adding a gray background for parameters that have a value. Activity items consist of only two parameters. The first one is a categorical parameter, which describes the binding type of the associated compound to the associated gene. As shown in Figure 6(b), we visualize the three categories, *activation*, *inhibition*, and *binding*, by icons that show an arrow pointing up, down, and a horizontal double arrow respectively. The second parameter describing the compound activity is the numerical AC50 value, which is encoded by a horizontal bar.

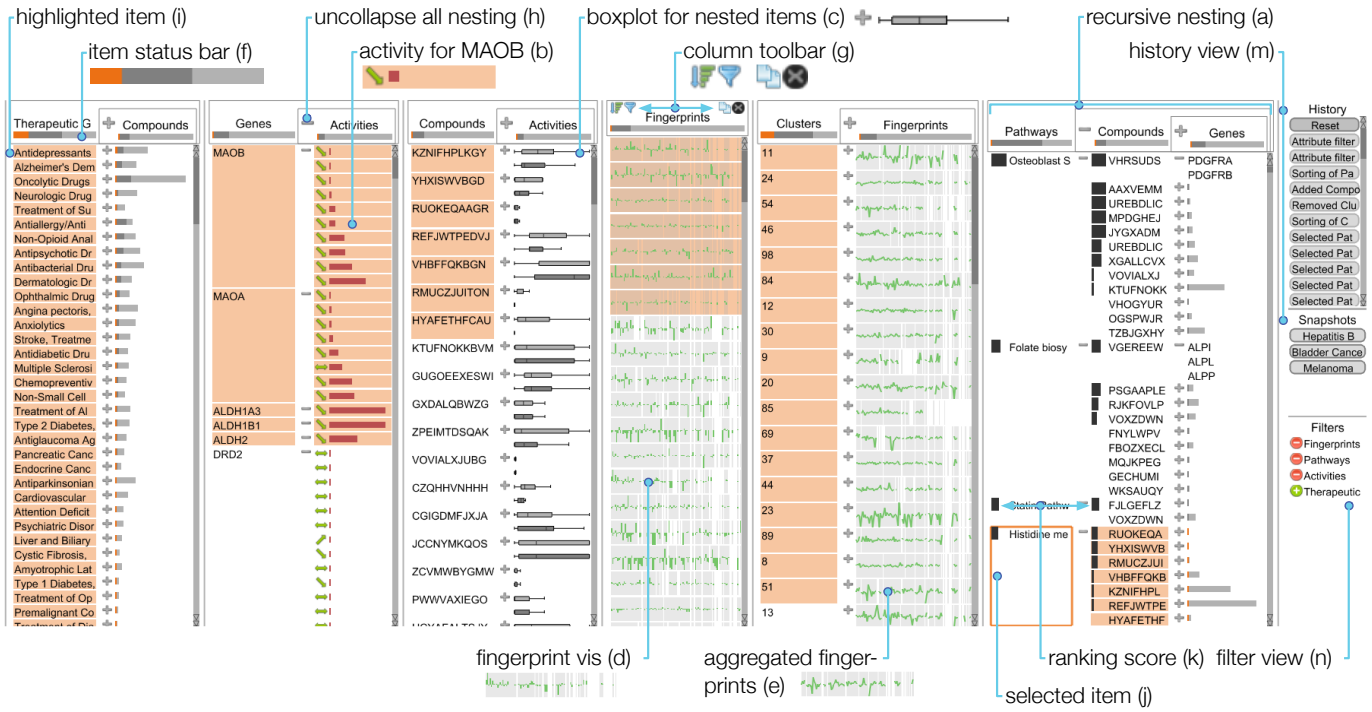


Fig. 6. Overview of visual components in the relationship view. Several different data items are listed in columns. (a) These columns can be nested recursively. The visual representations differ depending on the item type. By default items are represented as text. (b) Attributes of activity items are shown with bars and icons. (c) Box plots are used to summarize nested activity items. (d) Numerical fingerprint vectors show their data in bar charts. (e) Line charts summarize the median values of the fingerprints in child columns. (f) Composite bars in each column header show the number of all items (light gray), items not filtered out (dark gray), and selected items (orange) of a column. (g) The column toolbar is shown on demand and contains buttons to sort, filter, duplicate, and remove the column. (h) The items of a child column can be collapsed and expanded using the buttons in the column headers and next to the items. Highlighted items (i) indicate relationships to the selection source (j). (k) Columns that are ranked by enrichment scores represent the scores as bars right next to the items. (m) The history view records every analysis step and also shows snapshots taken. (n) The filter view displays the sequence of currently applied filters.

Nesting is a crucial concept in ConTour. To summarize nested items, we employ different encodings. The default summary representation that is available for any type of item provides an overview of children by indicating the number of child items using three bars that are drawn on top of each other. The light gray bar indicates the total number of children of a parent item, not considering any filters. The dark gray bar takes filters into account, indicating the number of children that will be shown if the user expands the summary representation. Finally, the orange bar indicates the number of children that are currently highlighted by selection. As illustrated in Figure 6(f), the same representation is used to give an overview of the items in each column. As activity data is tightly coupled with genes and compounds, they are usually nested within these columns. Our collaborators emphasized that it is important to enable analysts to get a quick overview of all activity values associated with a gene or compound. To address this, we provide an additional summary view that encodes the activity value distribution of child items using two box plots, as illustrated in Figure 6(c). The box plots drawn in light gray show the distribution for all child items, not considering any filters, whereas the dark gray box plots show the distribution for only those child items that were not filtered out. As clusters are based on fingerprints, they are typically nested within the cluster column. In order to represent the characteristics of a cluster, the fingerprints are aggregated into a line plot that encodes the median values for the fingerprints' parameters, as shown in Figure 6(e). We decided to employ a line plot instead of a bar chart, to make the summary representation easy to distinguish from the representation of individual fingerprints.

7.2 Detail views

As it is not possible to show all data associated with several item types in the relationship view, we provide a number of detail views to make this information accessible. As previously discussed, all detail views are tightly linked with the relationship view and also with each other. Thus, selections or filter operations are propagated to all other views.

Pathway view. The pathway view, shown in Figure 1, displays a selected pathway and its contained genes using a texture from one of the supported pathway databases. In addition, the pathway view also displays compounds that interact with at least one of its genes, as well as the fingerprint clusters the compounds are associated with. The pathway view is designed to support two analysis goals: identify which compounds interact with which genes within their cellular context, and identify which compounds and clusters are specific to a pathway. As previously mentioned, specificity is an important quality measure for the domain experts. The more specific a cluster is to a pathway, i.e., the more compounds of the cluster interact with the pathway, the more likely the compounds-pathway interaction is biologically relevant. Clusters are encoded as bars on the left and right of the pathway. The height of the bar encodes how many compounds of the cluster interact with the pathway. Its saturation indicates how specific a cluster is—highly specific clusters are dark blue, while unspecific clusters are white. Next to the clusters, smaller rectangles represent the compounds. By hovering or selecting a compound or cluster, all interaction partners within the pathway are highlighted, enabling analysts to identify the exact binding partners of each compound. The nodes in the pathway are shaded in yellow if no compound interacts with them. Nodes with a white shading bind to one compound, whereas saturated purple nodes bind to many compounds. The compounds optionally adhere to the system-wide filters, which allows the domain expert to quickly asses the relevance of a pathway for the remaining items.

Compound view. Being able to access the chemical properties of compounds is important to our collaborators as it provides crucial information when reasoning about, for example, why compounds fall

into the same cluster, or why they bind to similar targets. To realize this, the compound view, which is shown in Figure 1, displays the chemical structures for multiple compounds together with their names.

Parallel coordinates view. Analysts can make use of the parallel coordinates view to visualize any kind of multi-dimensional data. In the context of the available data, only activity data and fingerprint data fall into this category. As activity data is already displayed in full detail in the relationship view, the parallel coordinates view is mainly used to display fingerprint data. By default the view shows all fingerprints, but can be toggled to respect applied filters.

7.3 Support views

Two support views provide orientation and more flexibility during the data exploration.

History view. Every step taken in the exploration of items and their relationships is based on decisions made by the analyst. However, in some cases, the path taken might lead to a dead end, or the analyst just wants to explore the data in multiple directions without starting the analysis from scratch. As shown in Figure 6(m), we provide a history view to address this issue. The history view records every step taken during the analysis and allows the analyst to go back and forth within the analysis path as desired. Each step taken adds a new element to the history view; information about the step is shown on demand as a tooltip. Selecting an element reverts the system to the state when the element was recorded. In addition, analysts can take snapshots of the current state, which they can return to at any time.

Filter view. Filtering is an operation that is executed very frequently. However, keeping track of filters without support is hard. Therefore, ConTour tracks all applied filters in the filter view. As illustrated in Figure 6(n), every filter is represented by an element, which displays the name of the item set the filter was applied on. We use two symbols to indicate whether the data space was reduced by the filter, or items were added. A more detailed description of the filters are shown in a tooltip. Filters can be removed from the filter view on demand.

7.4 Enrichment score

One crucial task of our collaborators is judging how specific two types of items are related considering a third item type. For example, they want to know what clusters show an enrichment in compounds that modulate a specific pathway. Abstracted to general set terms, we want to know for an item i of set I (clusters) the enrichment of items of set K (compounds) that reach one item j in set J (pathways). We define an enrichment score as follows: Let K_i be the set of items in item set K that are related to i and K_j be the set of items in K that are related to j. For a pair of items (i, j), we calculate the enrichment score $s_{i,j}(K)$ by

$$s_{i,j}(K) = \frac{|K_i \bigcap K_j|/|K_j|}{|K_i|/|K|} \quad (1)$$

This is also illustrated in Figure 7. The numerator of this term describes how specific j is related to i via K. To account for the fact that items in I that are related to many items in K are more likely to also have common items with items in J, we divide by the given denominator. As small overlaps of one or two items were generally not interesting for our collaborators, we include a threshold for the minimum number of common items. The score is calculated for all pairs (i, j), which in turn can be used to rank columns in the relationship view. However, a column only shows items of a single type, although the score is defined for pairs. Therefore, we use the maximum score, given by $r_{i,K,J} = \max_{\forall j \in J}(s_{i,j}(K))$ to determine the rank of every item i in its column. We display this score as a horizontal bar next to the item. To see the item pairs, the paired columns can be nested, as shown in Figure 6. While the parent items show the maximum score, the child items indicate the scores achieved with their parent.

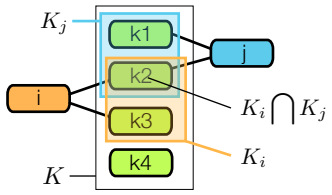


Fig. 7. Enrichment for i in j via K: $s_{i,j}(K) = (1/2)/(2/4) = 1$

7.5 Implementation

The ConTour visualization technique is a plugin for *Caleydo*[3], an open-source data visualization framework. Caleydo is implemented in Java and uses OpenGL/JOGL for rendering. The chemical compound structures are rendered using the *Chemistry Development Kit (CDK)* [25], an open-source Java library for structural chemo- and bioinformatics. The source code is available on GitHub[4]. ConTour will be part of the next major release of Caleydo.

8 Case Studies

ConTour is the result of a user-centered development process [20], which included regular meetings with our collaborators to iteratively develop and refine the system according to their needs. Together with our collaborators, who are chemical biologists, we conducted case studies to illustrate the applicability of ConTour on real-world problems. The overall goal of these case studies was to find out whether the biological fingerprints that were used as descriptors for compounds are able to detect meaningful biological similarities between compounds and reflect their effect on the cell and, ultimately, on the organism as a whole. If they prove to capture a compound's biological actions in a comprehensive manner, they can be used as a connecting module to identify relationships between compounds, targets, pathways, and diseases. In the following, we describe how our collaborators used ConTour to explore heterogeneous pharmaceutical and biological data and report on the gained insights.

8.1 Investigating phospodiesterase 4 inhibitors and their cluster neighbors

A straightforward way to explore the ability of the used descriptors to group compounds in a biologically meaningful way is to analyze the fingerprints of compounds that are known to modulate the same protein target. Therefore, the expert started by focusing on a particular protein target, the enzyme phosphodiesterase 4 (PDE4), which is represented by multiple different enzyme subtypes (PDE4A-D) in the dataset. She added a selection-based filter to limit all displayed items to those related to PDE4A-D. By applying an attribute filter to the activity data, our collaborator set an upper threshold for AC50 values of one μM, which resulted in ten different compounds that inhibit PDE4. Encouragingly, the fingerprints of four of the ten compounds belonged to cluster 56, proving that their shared target activity was reflected by similar fingerprint activity patterns. She was then interested what other compounds were found in Cluster 56. Therefore, she added all items related to Cluster 56. Overall, the cluster consisted of ten compounds. By ranking the protein targets by their enrichment of Cluster 56, she learned that two of the newly added compounds bind to beta-adrenergic receptors (ADRB1, ADRB2), which are evolutionary unrelated to PDE4. Also, when she displayed the compound structures, she saw that these compounds were structurally very distinct from the PDE4 inhibitors. At first glance, it seemed surprising that structurally diverse compounds binding to different proteins have similar biological fingerprints and cluster together. However, this observation became better understandable when she integrated the therapeutic group column into the analysis. As shown in Figure 8, five compounds from Cluster 56, among them modulators of both PDE4

[3] http://www.caleydo.org
[4] https://github.com/Caleydo/

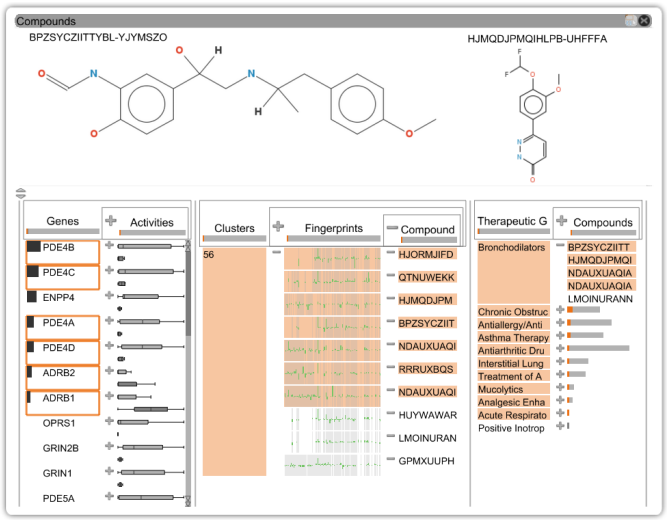


Fig. 8. The gene column on the left shows protein targets ranked by the enrichment of Cluster 56. The selected (orange frame) enzyme subtypes of phosphodiesterase 4 (PDE4A-D) and beta-adrenergic receptors (ADRB1, ADRB2) rank among the top. The highlights in the compound column, which is nested within the fingerprint and cluster columns, reveal that compounds binding to these targets fall into the same fingerprint cluster, Cluster 56. Differences in the chemical structures between compounds binding to ADRB1-2 and those binding to PDE4A-D can be seen in the detail view. When looking at the compound column nested within the therapeutic groups column, the highlights indicate that four out of the seven compounds that modulate ADRB or PDE4 map to the group *bronchodilators*. This group refers to compounds that are used for the treatment of breathing difficulties. Thus, the common bronchodilatory effects of compounds modulating these two target classes might be the reason why their fingerprints cluster together.

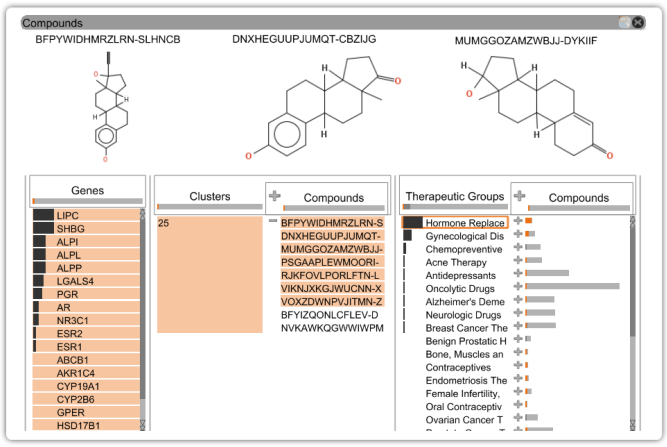


Fig. 9. To find out whether similarities in the fingerprints of compounds correlated with the compounds' classification in therapeutic groups, an enrichment score was used to calculate the enrichment of therapeutic groups in the fingerprint clusters. The highest score was achieved by Cluster 25 and the group *hormone replacement therapy* (selected item with orange frame), which was thus ranked on top of the therapeutic group column. As indicated by the highlight, this group and Cluster 25 have 7 compounds in common. The detail view of these compounds reveals that they are also structurally similar.

and adrenergic receptors, are mapped to the group *bronchodilators*, which refers to compounds that are used for the treatment of breathing difficulties. Indeed, it is known that both phosphodiesterases and adrenergic receptors are protein targets that are involved in airway diseases [17, 1]. Hence, as hoped for by the expert, the biological descriptor, i.e., the clustered fingerprints were able to find therapeutic effect similarities between compounds. In summary, ConTour helped to reveal complex relationships between compounds, their evolutionary unrelated targets, and diseases; and supported the hypothesis that compounds in Cluster 56 induced therapeutically relevant, similar cellular phenotypes. Based on this finding, it would be straightforward to suggest experiments that test the other compounds in the cluster for their bronchodilatory effects and their use for treating airway diseases.

8.2 Exploring overrepresentation of compounds with the same therapeutic effect

The previous case study has shown that compounds used for the treatment of airway diseases cluster together. To find out whether this observation could also be made for other therapeutic groups, our collaborator defined an enrichment score to systematically calculate the enrichment of therapeutic groups in the fingerprint clusters. By setting a minimum overlap threshold, she constrained the results to only contain therapeutic groups that are related to at least three different compounds in a cluster. The ranking of therapeutic groups by the enrichment score revealed that many different therapeutic groups were overrepresented in individual clusters. For example, seven therapeutic groups showed a strong, more than 20-fold enrichment in at least one cluster. Thus, the observation made for airway diseases was not a singular case but clustering of compounds that are administered for the same therapeutic indication is generally observed across the data set.

The highest enrichment score (87.9) was obtained for the group *hormone replacement therapy*. Of eight compounds mapping to this therapeutic group, seven fell into Cluster 25 and were grouped together with four more compounds of which two were annotated for hormone-related therapies. In accordance with the therapeutic role of the compounds, known targets that were enriched in this cluster were progesterone, androgen, and estrogen (sex hormone) receptors. Furthermore, displaying chemical structures revealed that all eleven compounds were chemically very similar steroids (see Figure 9). Hence, in this case, the chemical similarity of compounds correlates very well with their biological similarity.

8.3 Explaining cluster compositions by pathways

Finally, our collaborator wanted to find out whether compounds modulating different targets in the same biological pathway had similar biological fingerprints and thus cluster together. This could be expected considering that all targets in a pathway work together and that modulations of different targets in the same signaling cascade of a pathway should be propagated, ultimately leading to similar cellular phenotypes. Our collaborator started the analysis by selecting the important *JAK-STAT signaling pathway* shown in Figure 10. This pathway is evolutionary conserved across many different species and plays a role in the regulation of the immune system. When she opened the pathway in the detail view, she found four clusters that contained at least two compounds mapping onto the pathway. Interestingly, for all four clusters, it could be observed that their compounds modulated different targets in the pathway. The most prominent example was Cluster 69, in which six of overall 17 compounds in the cluster modulated one or more targets in the pathway. Of the six compounds, four inhibited different members of a protein family known as kinases, i.e., Janus kinases (JAKs) and the Pim-1 kinase, in the pathway. The other two compounds modulated the pathway by inhibiting the signal transducer STAT and the protein BCL that regulates transcription, i.e., gene expression. The collaborator confirmed that, without pathway information, clustering of compounds with diverse target activities is often difficult to reconcile. However, ConTour makes it possible to explain similar biological fingerprints of compounds with diverse target activities by showing that their targets act in a concerted manner in the cell by contributing to the same signaling event.

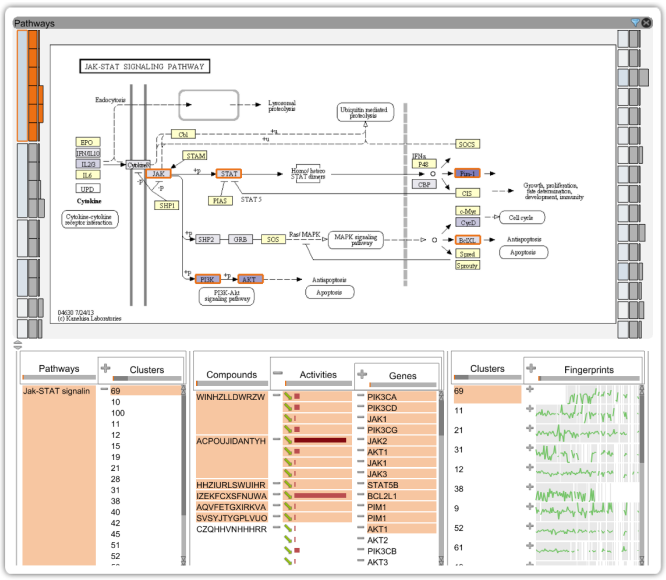


Fig. 10. The detail view shows the JAK-STAT signaling pathway. Selecting the block of Cluster 69 on the top left shows that the compounds of this cluster map to several different targets (graph nodes highlighted with an orange frame) in the pathway. The selection also highlights the genes and compounds in the relationship view. The (recursive) nesting of the gene column within the compound column displays the exact compound-target mappings and reveals that the compounds bind to different targets. This observation suggests that despite the compounds binding to different targets, their fingerprints clustered together because their targets are part of the same pathway.

9 Discussion

By using ConTour our collaborators gained several insights that support the hypothesis that biological fingerprints indeed reflect similarities among compounds and their biological effects on both the target protein and pathway level. By observing our collaborator during the analyses, we found that she picked up ConTour's concepts quickly. She used selections to identify relationships of individual items and filters to focus on the data of interest. The expert excessively used nesting, as she argued it helped understanding relationships of multiple items much better when she had to identify relationships across columns. To our surprise, she used nesting not in a static "set up once" approach, but constantly refined, removed, and added nestings to answer specific questions. We observed that she continuously used a combination of reasoning based on visualization and refinement using analytical processes and queries. For example, she relied on ranking by various scores and filtering to identify interesting items, but then refined her queries and adjusted her next steps based on the visual representations.

The combination of the query-driven relationship view and the various detail views proved highly valuable to our collaborators. Especially the compound detail view was frequently used to reason about whether observations made for compounds may be caused by their chemical properties, but also the pathway view was employed to contextualize the findings. In summary, the case studies confirm that the interplay of ConTour's building blocks is an effective approach for exploring relationships in drug discovery.

Scalability As a tool for exploring multi-relational data, ConTour needs to scale with respect to the number of columns, the number of items inside the columns, the number of nested columns, and the number of detail views it can handle effectively. In the case studies we demonstrated that ConTour can comfortably deal with about a dozen columns. Depending on the kind of data, we observed a limit of about 20 columns on a full-HD display. To even further increase the upper limit for the number of displayable columns, it would be possible to add a level-of-detail approach that lets the user manage larger number of columns. In terms of scalability of items, we have successfully worked with multiple columns containing up to 14,000 items. However, conceptually the column-based approach in combination with our task-dependent sorting of items scales to datasets with many more items. Regarding the recursive nesting of columns we found that more than four levels of nesting are rarely used in practice. Our current implementation supports about as many levels of recursions as columns for 1:1 relationships between datasets (i.e., about 20), while this number shrinks to about five for the n:m case. The number of detail views that can be shown simultaneously depends on the nature and size of the data and the used visualization. For compounds, we observed up to 8 simultaneously used views, while pathways were limited to one.

10 Conclusion

In this paper we introduced ConTour, a visual analysis system designed to facilitate the exploration of relationships in large cohorts of biological and pharmaceutical data. The system was developed in close collaboration with a team of chemical biologists at the pharmaceutical company Novartis. The main interface of the system displays entities of diverse datasets in a simple yet effective column-based layout. By combining sorting, filtering, and ranking strategies, analysts can drill down to the relevant items. Diverse interaction techniques allow users to browse the relationships of items in an environment of tightly interlinked views. Although the analysis of such complex interrelated data still poses many challenges, our collaborators confirmed that ConTour is an efficient tool to analyze multi-relational data for drug discovery.

In the future, we plan to investigate how our analysis approach can be applied to data graphs that contain cycles, as this property imposes challenges in terms of path ambiguities when resolving relationships in the graph traversal. A related problem is to also consider relationships between items of the same entity, for instance, relationships between genes in a pathway. To remedy this, we plan to merge the data graph with the complex cellular interaction networks that are captured in the pathways.

Our collaboration partners noted that they possess many similarly structured datasets that they plan to analyze using ConTour in the near future. Although we have designed ConTour specifically for the requirements in drug discovery, we argue that the approach can also be applied to both other biological and non-biological domains if the data is structured similarly. As the data structure closely resembles those of data stored in relational databases, we expect a broad applicability. For example, we are currently working with cancer researchers who use ConTour for the combined analysis of annotated tumor tissue images and mutation data.

Acknowledgments

The authors wish to thank Felix Reisen, Mark Borowsky, and the anonymous reviewers for their valuable input and feedback. This work was supported in part by the Austrian Science Fund (P22902, J 3437-N15), the Province of Styria HTI (A3-22.M-5/2012-21 "Tumor Heterogeneity") and the Air Force Research Laboratory and DARPA grant FA8750-12-C-0300. Anne Mai Wassermann is the recipient of a NIBR Presidential Postdoctoral Fellowship.

References

[1] K. H. Abbott-Banner and C. P. Page. Dual PDE3/4 and PDE4 inhibitors: Novel treatments for COPD and other inflammatory airway diseases. *Basic & Clinical Pharmacology & Toxicology*, 2014.

[2] O. M. Becker. Representing protein and peptide structures with parallel-coordinates. *Journal of Computational Chemistry*, 18(15):1893–1902, 1997.

[3] M. Dork, N. Riche, G. Ramos, and S. Dumais. PivotPaths: Strolling through faceted information spaces. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '12)*, 18(12):2709–2718, 2012.

[4] C. Dunne, N. Henry Riche, B. Lee, R. Metoyer, and G. Robertson. GraphTrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. In *Proceedings of the ACM SIGCHI*

Conference on Human Factors in Computing Systems (CHI '12), page 1663–1672. ACM, 2012.

[5] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(Database issue):D1100–1107, 2012.

[6] S. Ghani, B. C. Kwon, S. Lee, J. S. Yi, and N. Elmqvist. Visual analytics for multimodal social network analysis: A design study with social scientists. *IEEE Transactions on Visualization and Computer Graphics (VAST '13)*, 19(12):2032–2041, 2013.

[7] C. Görg, H. Tipney, K. Verspoor, W. A. B. Jr, K. B. Cohen, J. Stasko, and L. E. Hunter. Visualization and language processing for supporting analysis across the biomedical literature. In R. Setchi, I. Jordanov, R. J. Howlett, and L. C. Jain, editors, *Knowledge-Based and Intelligent Information and Engineering Systems*, number 6279 in Lecture Notes in Computer Science, pages 420–429. Springer, 2010.

[8] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42(D1):D199–D205, 2014.

[9] T. Kelder, M. P. v. Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico. WikiPathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(D1):D1301–D1307, Nov. 2011.

[10] S. Konecni, J. Zhou, and G. Grinstein. A visual analytics model applied to lead generation library design in drug discovery. In *13th International Conference on Information Visualisation*, pages 345–352, 2009.

[11] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(Database issue):D1091–1097, Jan. 2014.

[12] M. D. Lieberman, S. Taheri, H. Guo, F. Mir-Rashed, I. Yahav, A. Aris, and B. Shneiderman. Visual exploration across biomedical databases. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):536–550, 2010.

[13] E. Lounkine, P. Kutchukian, P. Petrone, J. W. Davies, and M. Glick. Chemotography for multi-target SAR analysis in the context of biological pathways. *Bioorganic & Medicinal Chemistry*, 20(18):5416–5427, 2012.

[14] J. Mestres, E. Gregori-Puigjané, S. Valverde, and R. V. Solé. Data completeness—the achilles heel of drug-target networks. *Nature Biotechnology*, 26(9):983–984, 2008.

[15] C. Partl, A. Lex, M. Streit, D. Kalkofen, K. Kashofer, and D. Schmalstieg. enRoute: Dynamic path extraction from biological pathway maps for exploring heterogeneous experimental datasets. *BMC Bioinformatics*, 14(Suppl 19):S3, 2013.

[16] P. M. Petrone, B. Simms, F. Nigsch, E. Lounkine, P. Kutchukian, A. Cornett, Z. Deng, J. W. Davies, J. L. Jenkins, and M. Glick. Rethinking molecular similarity: Comparing compounds on the basis of biological activity. *ACS Chemical Biology*, 7(8):1399–1409, 2012.

[17] B. J. Proskocil and A. D. Fryer. Beta2-agonist and anticholinergic drugs in the treatment of lung disease. *Proceedings of the American Thoracic Society*, 2(4):305–310; discussion 311–312, 2005.

[18] R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*, pages 318–322. ACM, 1994.

[19] H.-J. Schulz, M. John, A. Unger, and H. Schumann. Visual analysis of bipartite biological networks. In *Proceedings of the Eurographics Workshop on Visual Computing for Biomedicine (VCBM '08)*, pages 135–142, 2008.

[20] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431 –2440, Dec. 2012.

[21] B. Shneiderman and A. Aris. Network visualization by semantic substrates. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '06)*, 12(5):733–740, 2006.

[22] M. Spenke and C. Beilken. InfoZoom - analysing formula one racing results with an interactive data mining and visualisation tool. In *Proceedings of the International Conference on Data Mining*, page 455–464, 2000.

[23] M. Spenke, C. Beilken, and T. Berlage. FOCUS: The interactive table for product comparison and selection. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '96)*, page 41–50. ACM, 1996.

[24] J. Stasko, C. Gorg, Z. Liu, and K. Singhal. Jigsaw: Supporting investigative analysis through interactive visualization. In *Proceedings of the IEEE Symposium on Visual Analytics in Science and Technology (VAST '07)*, pages 131–138. IEEE, 2007.

[25] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The chemistry development kit (CDK): An open-source java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2):493–500, 2003.

[26] K. Strebhardt and A. Ullrich. Paul ehrlich's magic bullet concept: 100 years of progress. *Nature reviews. Cancer*, 8(6):473–480, 2008.

[27] H. Strobelt, E. Bertini, J. Braun, O. Deussen, U. Groth, T. U. Mayer, and D. Merhof. HiTSEE KNIME: a visualization tool for hit selection and analysis in high-throughput screening experiments for the KNIME platform. *BMC Bioinformatics*, 13(Suppl 8):S4, 2012.

[28] A. M. Wassermann, E. Lounkine, and M. Glick. Bioturbo similarity searching: combining chemical and biological similarity to discover structurally diverse bioactive molecules. *Journal of chemical information and modeling*, 53(3):692–703, 2013.

[29] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*, page 401–408. ACM, 2003.

[30] J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan. Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE Transactions on Visualization and Computer Graphics (VAST '13)*, 19(12):2080–2089, 2013.